12: Solutions
Stat 120 | Fall 2025

Prof Amanda Luby

```
library(tidyverse)
library(broom)
library(patchwork)
library(CarletonStats)
```

The data set Pew.csv contains part of a survey conducted by the Pew Research Center in January 2014. One of the questions they asked was, "Overall, when you add up all the advantages and disadvantages of the internet, would you say that the internet has been mostly a good thing, a bad thing, or some of both?"

The variable `values` codes the response as "good" if the respondent said the internet has been a good thing and "bad" otherwise (this includes "a bad thing" and "some of both").

Let's see if this differs based on whether the respondent is 50 years or older or not.

```
Pew <- read.csv("http://math.carleton.edu/Stat120/RLabManual/Pew.csv")

Pew <- Pew |>
  mutate(
    over = ifelse(age >=50, "over50", "under50"),
    over2 = ifelse(age >=50, 1, 0)
  )
```

1. Create an appropriate visualization of `values` conditioned on `over`. Does it appear that the proportion of over 50 year olds that said the internet has been a good thing is approximately the same as the under 50's?

> **Solution**
>
> You can tell they are not exactly equal, but the percentage who say it's been a good thing appear pretty similar across the two age groups

```
ggplot(Pew, aes(y = over, fill = values)) +
  geom_bar(position = "fill")
```
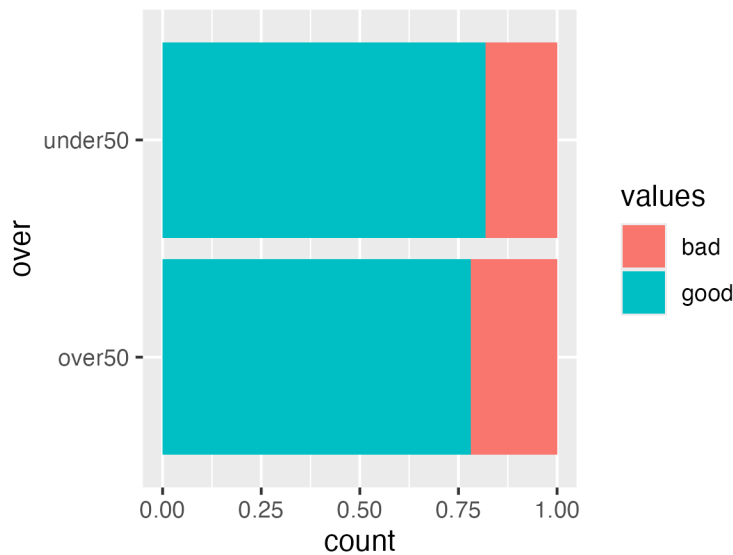

2. Compute the exact proportions for each age group.

Solution

```
table(Pew$over, Pew$values) |>
  proportions(2)
```

```
              bad      good
  over50   0.5714286 0.5125899
  under50 0.4285714 0.4874101
```

3. In order to use the **boot()** command, the response must be coded as a binary variable of 0's and 1's (instead of the names of the categories). Pull up the "spreadsheet view" of the data and confirm that **values** and **values2** contain the same information stored in a different form.

Solution

Confirmed

Since the values are 0's and 1's, when you compute the mean, you will be computing an expression of the form

$$\frac{1 + 1 + 0 + 1 + 0 + \ldots + 0}{n}$$

which is equivalent to the proportion of 1's. This means that we can treat proportions as means. Yay!

4. Create a bootstrap distribution of the difference in proportions and give a sentence interpreting the 95% percentile interval. Use the variable `values2`, which is behavior recoded so that 1 = "good", 0 = "bad".

Solution

I am 95% confident that the true difference in proportions between the two age groups is between -0.1 and 0.02.

```
boot(values2 ~ over, data = Pew)
```

```
    ** Bootstrap interval for difference of mean

 Observed difference of mean : over50 - under50 = -0.03791
 Mean of bootstrap distribution: -0.03736
 Standard error of bootstrap distribution: 0.03051

 Bootstrap percentile interval
       2.5%        97.5%
-0.09797624  0.02272069

        *--------------*
```
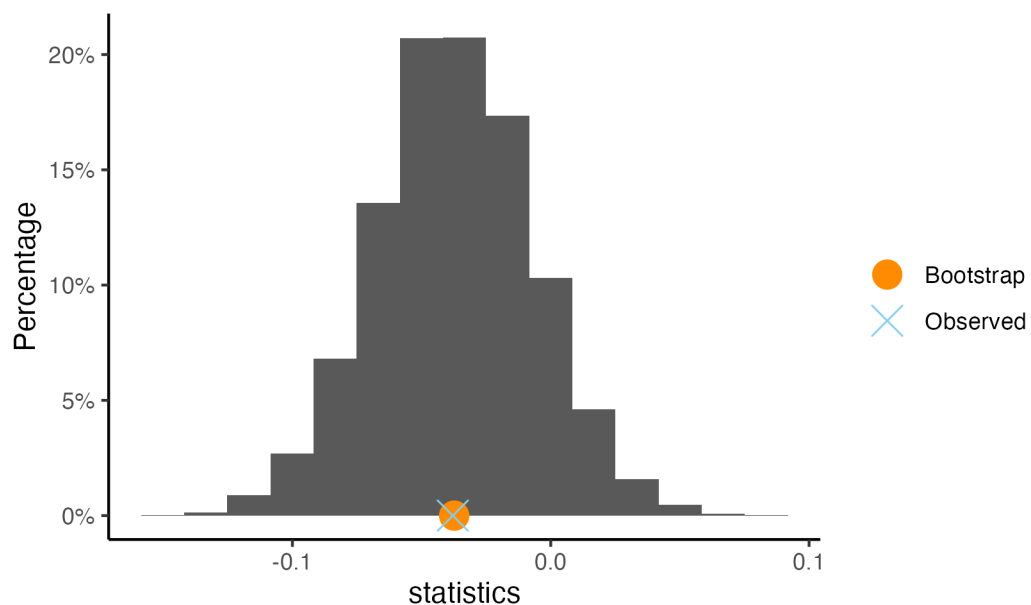

Bootstrap distribution

---

5. Compute the 95% confidence interval based on the 95% rule. How does it compare to

the percentile interval?

Solution

$$-.0379 \pm 2 \times 0.03 = [-0.0979, 0.0221]$$

It is extremely similar.